INFERENCIA ESTADÍSTICA

Notas

Indice

1. Objetivos

- Comprender los fundamentos lógico-matemáticos de la inferencia estadística;
- Saber plantear, resolver e interpretar problemas de intervalos de confianza;
- Comprender las distintas fases de un contraste de hipótesis;
- Saber plantear, resolver e interpretar problemas de contraste de hipótesis;
- Analizar los errores que pueden cometerse en un contraste de hipótesis;
- Enjuiciar la correcta aplicación de las pruebas estadísticas en situaciones de investigación concretas;
- Saber interpretar el resultado del valor p a la hora de tomar decisiones en los contrastes de hipótesis.

2. Inferencia estadística

La inferencia estadística es aquella rama de la estadística mediante la cual se trata de sacar conclusiones de una población en estudio, a partir de la información que proporciona una muestra representativa de la misma. También es denominada “estadística inductiva” o “inferencia inductiva” ya que es un procedimiento para generar nuevo conocimiento científico.

La muestra se obtiene por observación o experimentación. La necesidad de obtener un subconjunto reducido de la población es obvia si se tiene en cuenta los costes económicos de la experimentación o el hecho de que muchos de los métodos de medida son destructivos.

Toda inferencia inductiva exacta es imposible ya que se dispone de información parcial, sin embargo es posible realizar inferencias inseguras y medir el grado de inseguridad si el experimento se ha realizado de acuerdo con determinados principios. Uno de los propósitos de la inferencia estadística es el de conseguir técnicas para hacer inferencias inductivas y medir el grado de incertidumbre de tales inferencias. La medida de la incertidumbre se realiza en términos de probabilidad.

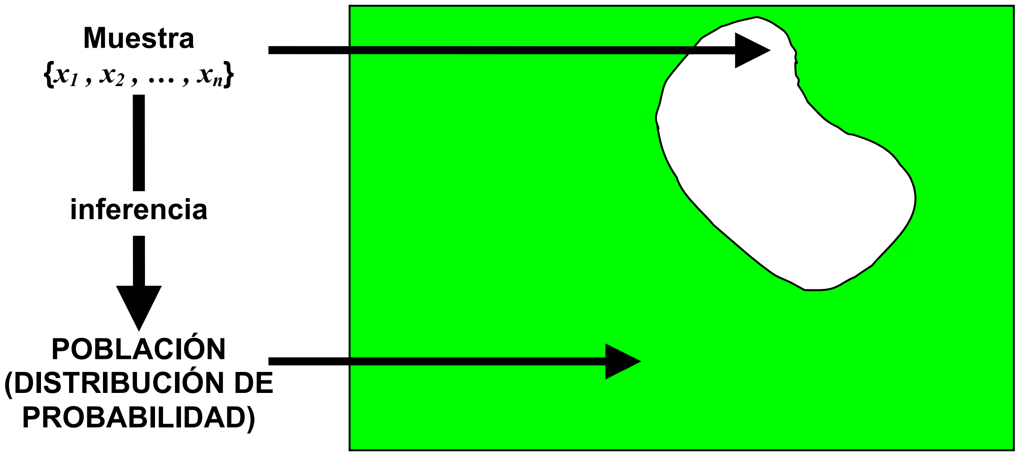

Figura 1. Esquema de inferencia estadística.

De acuerdo con el conocimiento sobre la distribución en la población, la inferencia estadística puede dividirse en:

- Inferencia paramétrica: se conoce la forma de la distribución (normal, binomial, Poisson, etc...) pero se desconocen sus parámetros. Se realizan inferencias sobre los parámetros desconocidos de la distribución conocida;
- Inferencia no paramétrica: forma y parámetros desconocidos. Se realizan inferencias sobre características que no tienen porque ser parámetros de una distribución conocida (mediana, estadísticos de orden).

De acuerdo con la forma en que se estudian los parámetros o características desconocidas, la inferencia puede dividirse en:

- Estimación: se intenta dar estimaciones de los parámetros desconocidos sin hacer hipótesis previas sobre posibles valores de los mismos:
    - Estimación puntual: un único valor para cada parámetro;
    - Estimación por intervalos: intervalo de valores probables para el parámetro.
- Contraste de hipótesis: se realizan hipótesis sobre los parámetros desconocidos y se desarrolla un procedimiento para comprobar la verosimilitud de la hipótesis planteada.

3. Intervalos de confianza

En los métodos de *estimación puntual* se utiliza una función de los valores de la muestra (estadístico) para dar la estimación del parámetro en estudio. Si en vez de esto, se utilizan dos funciones y se da el valor de dicho parámetro a partir del intervalo que tiene por extremos los valores de dichas funciones para una muestra, se dice que se está dando una *estimación por intervalos* del parámetro, o un *intervalo de confianza*.

En la construcción de estos intervalos, hay dos elementos fundamentales. La amplitud del intervalo que dará la precisión de la estimación, y que por lo tanto deberá ser la menor posible, y la probabilidad de que el intervalo contenga al verdadero valor del parámetro a estimar, que se llama nivel de confianza o coeficiente de confianza, y que deberá ser lo mayor posible. Está claro que se puede ganar en precisión a base de perder confianza en la estimación.

Para centrar ideas, véase el caso en el que se selecciona una muestra aleatoria simple de una población descrita por la función $f(x;\theta)$ dependiente del parámetro θ que pretendemos estimar. El problema se plantea como sigue: se fija un nivel de confianza, que se denota por $1-\alpha$, en donde $0<\alpha<1$, y se trata de determinar dos funciones

$$\theta_1(x_1, x_2, \ldots, x_n)$$
$$\theta_2(x_1, x_2, \ldots, x_n)$$

de forma que:

$$\Pr\left(\theta_1(x_1, x_2, \ldots, x_n) \le \theta \le \theta_2(x_1, x_2, \ldots, x_n)\right) = 1-\alpha$$

Al intervalo $\left[\theta_1(x_1, x_2, \ldots, x_n), \theta_2(x_1, x_2, \ldots, x_n)\right]$ se le llama intervalo de confianza de θ al nivel de confianza del $(1-\alpha)\times 100\%$. Es muy importante observar que sería un error afirmar que la probabilidad indicada anteriormente, es la probabilidad de que θ esté entre los números reales $\theta_1(x_1, x_2, \ldots, x_n)$ y $\theta_2(x_1, x_2, \ldots, x_n)$ puesto que θ no es una variable aleatoria sino un parámetro que tendrá un valor concreto, aunque sea desconocido. Las variables aleatorias son $\theta_1(x_1, x_2, \ldots, x_n)$ y $\theta_2(x_1, x_2, \ldots, x_n)$ al variar la muestra, luego la probabilidad anterior debe ser considerada como la probabilidad de que el intervalo aleatorio $\theta_1(x_1, x_2, \ldots, x_n)$a $\theta_2(x_1, x_2, \ldots, x_n)$ contenga el verdadero valor de θ. Dicho en términos de frecuencias esto significa que de cada 100 muestras aleatorias que se tomen, cabe esperar que el $(1-\alpha)\times 100\%$ de ellas contenga al verdadero valor de θ entre $\theta_1(x_1, x_2, \ldots, x_n)$ y $\theta_2(x_1, x_2, \ldots, x_n)$.

El esquema general para la estimación de un intervalo de confianza es:

$$\textit{estimador} \pm \textit{coeficiente de confiabilidad} \times \textit{error estándard}$$

El coeficiente de confiabilidad (z o t) indica entre más/menos cuántos errores estándar del estimador está $1-\alpha$ del área de la distribución muestral del estadístico.

La probabilidad de que los valores obtenidos a través del estimador por medio de un intervalo contenga el verdadero valor del parámetro que se pretende estimar de la población, es $1-\alpha$. La probabilidad de estimar $1-\alpha$ de la población se llama también "*coeficiente de confianza*" o "*probabilidad de acertar*". Los coeficientes de confianza más utilizados son: 0,90, 0,95 y 0,99.

La probabilidad α ("probabilidad de equivocarse") se divide en dos áreas en los extremos $\left(\alpha/2\right)$.

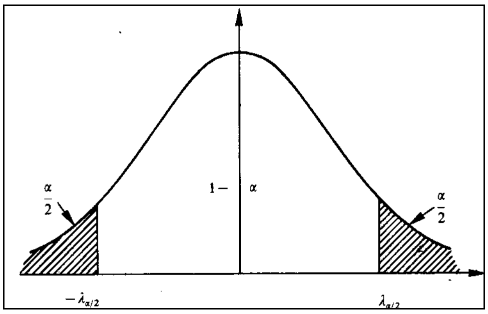

Figura 2.

Valores del coeficiente de confiabilidad si $n > 30(z)$.

Para $1-\alpha = 99\%$, $\alpha = 1\% = 0{,}01$

$z_{\alpha/2} = \pm 2{,}58$

Para $1-\alpha = 95\%$, $\alpha = 5\% = 0{,}05$

$z_{\alpha/2} = \pm 1{,}96$

Para $1-\alpha = 90\%$ $\alpha = 10\% = 0{,}10$

$z_{\alpha/2} = \pm 1{,}64$

A la mitad de la amplitud del intervalo de confianza se le llama *precisión del estimador.*

En todo intervalo de confianza hay un aspecto positivo y un aspecto negativo:

- El positivo, se ha usado una técnica que acierta en una alta proporción de casos;
- El negativo, se desconoce si en el caso concreto se ha acertado.

4.1. Intervalos de confianza para la media

Se supone que $\{x_1, x_2, \ldots, x_n\}$ es una muestra aleatoria simple y que $\overline{x}$ es la media muestral.

4.1.1.Población normal de varianza conocida.

En estas condiciones, la variable aleatoria $\overline{x}$ media muestral se distribuye según una distribución $N\left(\mu, \sigma^2/n\right)$, suponiendo que la población es $N\left(\mu, \sigma^2\right)$ con σ conocida. Esta propiedad permite asegurar que la variable tipificada $\dfrac{\overline{x}-\mu}{\sigma/\sqrt{n}}$ sigue una distribución $N(0,1)$, y por lo tanto, fijado el nivel de confianza $1-\alpha$, obtener el valor $z_{\alpha/2}$ tal que:

$$\Pr\left(\frac{\overline{x}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1-\alpha$$

Operando, resulta:

$$\begin{aligned} 1-\alpha &= \Pr\left(-z_{\alpha/2} \leq \frac{\overline{x}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) \\ &= \Pr\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \overline{x}-\mu \leq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \\ &= \Pr\left(-\overline{x}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\overline{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \\ &= \Pr\left(\overline{x}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \end{aligned}$$

con lo que las funciones θ_1 y θ_2 buscadas, en este caso, serán

$$\theta_1 = \overline{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$\theta_2 = \overline{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

y el intervalo de confianza para la media poblacional al nivel de confianza del (1 – α) 100% es:

$$\left[\overline{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \ , \ \overline{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

Se puede ganar en precisión de dos formas, bien perdiendo confianza lo que, en general, no interesa, o bien aumentando el tamaño n de la muestra seleccionada.

4.1.2.Población normal de varianza desconocida.

Sea ahora una población $N\left(\mu, \sigma^2\right)$ con σ desconocida. Es evidente que en este caso se han de utilizar estimadores que no dependan del valor σ. De aquí la importancia de las distribuciones que no dependen de este valor, como es el caso de la *t* de *Student*. En estas condiciones, la variable $\dfrac{\overline{x} - \mu}{\sigma / \sqrt{n-1}}$, es una *t-Student* con $n-1$ grados de libertad, en donde s es la desviación típica muestral. Esta propiedad permite, fijado el nivel de confianza $1-\alpha$, obtener el valor $t_{\alpha/2}$ tal que:

$$\Pr\left(\left| \frac{\overline{x} - \mu}{s / \sqrt{n-1}} \right| \leq t_{\alpha/2} \right) = 1-\alpha$$

Operando de igual forma que en el apartado anterior resulta:

$$\Pr\left(\overline{x} - t_{\alpha/2} \frac{s}{\sqrt{n-1}} \leq \mu \leq \overline{x} + t_{\alpha/2} \frac{s}{\sqrt{n-1}} \right) = 1-\alpha$$

y el intervalo de confianza en este caso es:

$$\left[\overline{x} - t_{\alpha/2} \frac{s}{\sqrt{n-1}} \ , \ \overline{x} + t_{\alpha/2} \frac{s}{\sqrt{n-1}} \right]$$

4.1.3 Intervalo de confianza para la media geométrica

El intervalo de confianza para la media geométrica viene dado por la expresión:

$$\left[e^{\left(\overline{x_{\log}} \pm z_{\alpha/2} es_{\log} \right)} \right]$$

Que es consecuencia de:

- Transformar los valores de la variable X en $\log\left(X\right)$;
- Hallar la media $\overline{x_{\log}}$ de los $\log\left(X\right)$;
- Hallar el error estándar $es_{\log}$ de la media de los $\log\left(X\right)$.

4.2. Intervalo de confianza para medianas y otros cuantiles

El intervalo de confianza para los cuantiles en general es:

$$\left[(n+1)q \;\pm\; z_{\alpha/2} \sqrt{n\,q\,(1-q)} \right]$$

Para un percentil 80, $q = 0{,}8$.

Para la mediana

$$\left[\frac{(n+1)}{2} \;\pm\; z_{\alpha/2} \frac{\sqrt{n}}{2} \right]$$

teniendo en cuenta que los valores corresponden al número de orden fijado por dicho intervalo y que ocupa cada dato de la variable cuando se ordenan de menor a mayor.

4.3. Intervalo de confianza para la proporción

El intervalo de confianza para una proporción p/q es:

$$\left[\hat{p} \;\pm\; z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

4.4. Intervalos de confianza para la varianza

Suponiendo que la población es normal de media μ y desviación típica σ, la variable aleatoria $\frac{n\,s^2}{\sigma^2}$ sigue una distribución χ^2 con $n-1$ grados de libertad. Esta propiedad, permite obtener dos valores $\chi^2_{\alpha/2}$ y $\chi^2_{1-\alpha/2}$ tales que:

$$\Pr\left(\chi^2_{1-\alpha/2} \le \frac{n\,s^2}{\sigma^2} \le \chi^2_{\alpha/2} \right) = 1-\alpha$$

luego:

$$\Pr\left(\frac{1}{\chi^2_{1-\alpha/2}} \le \frac{\sigma^2}{n\,s^2} \le \frac{1}{\chi^2_{\alpha/2}} \right) = 1-\alpha$$

con lo que resulta:

$$\Pr\left(\frac{n\,s^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{n\,s^2}{\chi^2_{1-\alpha/2}} \right) = 1-\alpha$$

por lo que el intervalo de confianza para la varianza poblacional, al nivel de confianza del (*1 – α) 100%*, viene dado por:

$$\left[\frac{n\,s^2}{\chi^2_{\alpha/2}} \;,\; \frac{n\,s^2}{\chi^2_{1-\alpha/2}} \right]$$

o de manera análoga:

$$\left[\frac{n-1\,\bar{s}^2}{\chi^2_{\alpha/2}} \;,\; \frac{n-1\,\bar{s}^2}{\chi^2_{1-\alpha/2}} \right] \qquad [1]$$

Obsérvese que, a diferencia de la estimación de la media por intervalos, el intervalo anterior no es único, es decir, que existen infinitos pares de valores $\chi^2_{\alpha/2}$ y $\chi^2_{1-\alpha/2}$ que verifican [1]. De entre estos pares de valores

deberían elegirse aquellos que den el intervalo de menor amplitud; sin embargo, al ser muy complicados los cálculos, se utilizan en la práctica los valores de $\chi^2_{\alpha/2}$ y $\chi^2_{1-\alpha/2}$ indicados en la figura:

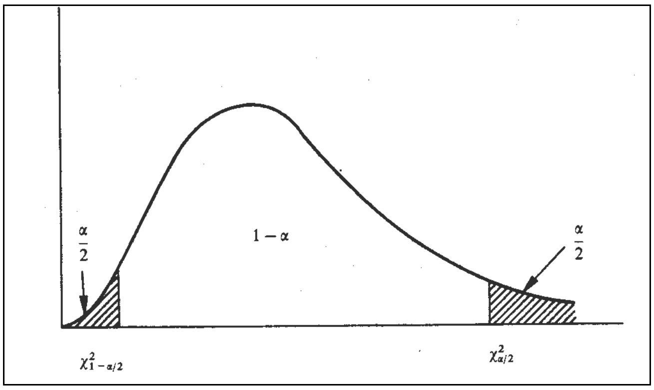

Figura 3.

4.5. Intervalo de confianza para la razón de varianzas

El intervalo de confianza para la razón s_1^2 / s_2^2 de varianzas es:

$$\left[\frac{s_1^2 / s_2^2}{F_{n_1-1,\, n_2-1,\, \alpha/2}} \; , \; \frac{s_1^2 / s_2^2}{F_{n_1-1,\, n_2-1,\, 1-\alpha/2}} \right]$$

4.6. Intervalo de confianza para el coeficiente de variación

El intervalo de confianza para el coeficiente de variación es:

$$\left[\nu \frac{CV}{\sqrt{j}} \; , \; \nu \frac{CV}{\sqrt{k}} \right]$$

donde $j = \chi^2_{\alpha/2}$;

$k = \chi^2_{1-\alpha/2}$; y

ν son los grados de libertad.

4.7. Intervalo de confianza para la diferencia de medias en poblaciones normales con varianza conocida

Sea $\{x_1, x_2, \ldots, x_n\}$ un muestreo aleatorio simple de $N(\mu_1, \sigma_1^2)$ y $\{y_1, y_2, \ldots, y_n\}$ uno de $N(\mu_2, \sigma_2^2)$. Ambas muestras son independientes. Supóngase que μ_1 y μ_2 son desconocidos y que σ_1 y σ_2 son conocidas. Se desea obtener un intervalo de confianza para $(\mu_1 - \mu_2)$ de nivel $1-\alpha$.

Se sabe que:

$$\bar{x} - \bar{y} \approx N\left(\mu_1-\mu_2\,,\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right) \Rightarrow \frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \approx N(0\,,1)$$

Existe un $z_{\alpha/2}$ que verifica:

$$\Pr\left(-z_{\alpha/2} \leq \frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \leq z_{\alpha/2}\right) = 1-\alpha$$

$$-z_{\alpha/2} \leq \frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \Rightarrow (\mu_1-\mu_2) \leq (\bar{x}-\bar{y})+z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

Despejando:

$$\frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \leq z_{\alpha/2} \Leftrightarrow (\mu_1-\mu_2) \geq (\bar{x}-\bar{y})-z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

Entonces:

$$\Pr\left(-z_{\alpha/2} \leq \frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \leq z_{\alpha/2}\right) = 1-\alpha \Leftrightarrow \Pr\left((\bar{x}-\bar{y})-z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \leq (\mu_1-\mu_2)\right)$$

Luego el intervalo de confianza de nivel $1-\alpha$ es:

$$\left[(\bar{x}-\bar{y}) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right]$$

4.8. Intervalo de confianza para la diferencia de medias de dos distribuciones normales, varianzas desconocidas pero iguales

Si $\bar{x}_1$, $\bar{x}_2$, s_1^2 y s_2^2 son las medias y las varianzas de dos muestras aleatorias de tamaño n_1 y n_2, respectivamente, tomadas de dos poblaciones normales e independientes con varianzas desconocidas pero iguales, el intervalo de confianza del $(1-\alpha)\times 100\,\%$ para la diferencia entre medias $(\mu_1-\mu_2)$ es:

$$\left[(\bar{x}_1-\bar{x}_2) \pm t\, s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right]$$

donde: $s_p^2 = \dfrac{s_1^2(n_1-1)+s_2^2(n_2-1)}{n_1+n_2-2}$ es el estimador combinado de la desviación típica común de la población con n_1+n_2-2 grados de libertad.

4.9. Intervalo de confianza para la diferencia de medias de dos distribuciones normales con varianzas desconocidas

El intervalo de confianza para la diferencia de medias $\mu_1 - \mu_2$ de dos distribuciones normales, varianzas desconocidas pero diferentes es:

$$\left[\left(\bar{x}_1 - \bar{x}_2 \right) \pm t \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]$$

donde: $\nu = \dfrac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2} \right)^2}{\dfrac{\left(\dfrac{s_1^2}{n_1} \right)^2}{n_1 - 1} + \dfrac{\left(\dfrac{s_2^2}{n_2} \right)^2}{n_2 - 1}}$ son los grados de libertad

4.10. Intervalo de confianza para la diferencia de medias con datos apareados

Sean

$$X \rightarrow N\left(\mu_X , \sigma_X\right)$$

$$Y \rightarrow N\left(\mu_Y , \sigma_Y\right)$$

De ambas se extraen dos muestras aleatorias:

$$\left\{x_1, x_2, \ldots, x_n\right\}$$

$$\left\{y_1, y_2, \ldots, y_n\right\}$$

Se define

$$D = X - Y$$

Que tiene la muestra asociada:

$$\left\{d_1 = x_1 - y_1 \ , \ d_2 = x_2 - y_2 \ , \ \ldots \ , \ d_n = x_n - y_n\right\}$$

El intervalo de confianza es:

$$\left[\bar{d} \ \pm \ t_{\alpha/2,\, n-1} \frac{s_d}{\sqrt{n}} \right]$$

4.11. Intervalo de confianza de la diferencia de proporciones

Recuérdese que para muestras de tamaño grande, el estimador $\hat{p}_1 - \hat{p}_2$ es aproximadamente normal con media $p_1 - p_2$ y varianza

$$\frac{p_1\left(1 - p_1\right)}{n_1} + \frac{p_2\left(1 - p_2\right)}{n_2}$$

El intervalo de confianza es, entonces:

$$\left[\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1\left(1 - \hat{p}_1\right)}{n_1} + \frac{\hat{p}_2\left(1 - \hat{p}_2\right)}{n_2}} \right]$$

4.12. Intervalo de confianza para el parámetro λ de una distribución de Poisson

El intervalo de confianza para el parámetro λ de una distribución de Poisson es:

$$\left[\left(\hat{\lambda} + \frac{z^2_{\alpha/2}}{2n}\right) \pm \frac{z_{\alpha/2}}{2}\sqrt{\left(\frac{z^2_{\alpha/2}}{n} + 4\hat{\lambda}\right)\frac{1}{n}}\right] \qquad \forall\ 10 \le \hat{\lambda} < 20$$

$$\left[\hat{\lambda} \pm z_{\alpha/2}\sqrt{\frac{\hat{\lambda}}{n}}\right] \qquad \forall\ \hat{\lambda} \ge 20$$

$$\left[\left(\max\left(0\,,\,\sqrt{\hat{\lambda}} - z_{\alpha/2}\frac{1}{2\sqrt{n}}\right)\right)^2 , \left(\sqrt{\hat{\lambda}} + z_{\alpha/2}\frac{1}{2\sqrt{n}}\right)^2\right]$$

donde: $n = \left(\frac{z_{\alpha/2}}{d}\right)^2 \lambda$

4.13. Intervalo de confianza para la diferencia de parámetros de dos distribuciones de Poisson independientes

El intervalo de confianza para la diferencia de parámetros $\lambda_1 - \lambda_2$ de dos distribuciones de Poisson independientes es:

$$\left[\left(\hat{\lambda}_1 - \hat{\lambda}_2\right) \pm z_{\alpha/2}\sqrt{\frac{\hat{\lambda}_1}{n_1} + \frac{\hat{\lambda}_2}{n_2}}\right]$$

5. Contrastes de hipótesis

5.1. Conceptos generales

Una *hipótesis estadística* es una afirmación que se hace acerca de una o varias características de una población.

Un *contraste de hipótesis* es un procedimiento para decidir si una hipótesis se acepta como válida o se rechaza. Su finalidad esencial consiste en demostrar o rechazar hipótesis científicas, mediante un razonamiento inductivo de tipo probabilístico.

Los métodos de contraste de hipótesis tienen como objetivo comprobar si determinado supuesto referido a un parámetro poblacional, o a parámetros análogos de dos o más poblaciones, es compatible con la evidencia empírica contenida en la muestra. Los supuestos que se establecen respecto a los parámetros se llaman hipótesis paramétricas. Para cualquier hipótesis paramétrica, el contraste se basa en establecer un criterio de decisión, que depende en cada caso de la naturaleza de la población, de la distribución de probabilidad del estimador de dicho parámetro y del control que se desea fijar a priori sobre la probabilidad de rechazar la hipótesis contrastada en el caso de ser ésta cierta.

Referente al contraste de hipótesis, se sabe que un problema es investigable cuando existen dos o más soluciones alternativas y se tienen dudas acerca de cual de ellas es la mejor. Esta situación permite formular una o más hipótesis de trabajo, ya que cada una de ellas destaca la conveniencia de una de las soluciones sobre las demás. Si el propósito es comprobar una teoría ella misma será la hipótesis del trabajo, pero es importante destacar que al formular dicha o dichas hipótesis no significa que ya esté resuelto el problema, al contrario, que la duda impulsa a comprobar la verdad o falsedad de cada una de ellas. La decisión final partirá de las decisiones previas de aceptar o rechazar las hipótesis de trabajo.

Ejemplo 1

Por ejemplo, supóngase que debe realizarse un estudio sobre la altura media de los habitantes de cierto pueblo de España. Antes de tomar una muestra, lo lógico es hacer la siguiente suposición *a priori*, (hipótesis que se desea contrastar y que se denota como H_0:

$$H_0 \;:\; \textit{la altura media no difiere de la del resto del país}$$

Al obtener una muestra de tamaño $n=8$, se podría encontrar ante uno de los siguientes casos:

$$\textit{caso } 1 \quad \textit{muestra} \;=\; \{1{,}50, 1{,}52, 1{,}48, 1{,}55, 1{,}60, 1{,}49, 1{,}55, 1{,}63\}$$
$$\textit{caso } 2 \quad \textit{muestra} \;=\; \{1{,}65, 1{,}80, 1{,}73, 1{,}52, 1{,}75, 1{,}65, 1{,}75, 1{,}78\}$$

Intuitivamente, en el caso 1 sería lógico suponer que salvo que la muestra obtenida sobre los habitantes del pueblo sea muy poco representativa, la hipótesis H_0 debe ser rechazada. En el caso 2 tal vez no se pueda afirmar con rotundidad que la hipótesis H_0 sea cierta, sin embargo no se podría descartar y se admitirá por una cuestión de simplicidad.

En todo contraste intervienen dos hipótesis. La H_0 es aquella que recoge el supuesto de que el parámetro toma un valor determinado y es la que soporta la carga de la prueba. Es la que se acepta como verdadera, la que se pretende rechazar y la que afirma que no existe diferencia entre dos poblaciones, o entre los parámetros del estudio. La decisión de rechazar la H_0, que en principio se considera cierta, está en función de que sea o no compatible con la evidencia empírica contenida en la muestra. El contraste clásico permite controlar a priori la probabilidad de cometer el error de rechazar la H_0 siendo ésta cierta; dicha probabilidad se llama nivel de significación del contraste (α) y suele fijarse en el 1 %, 5 % ó 10 %.

La proposición contraria a la H_0 recibe el nombre de hipótesis alternativa (H_1) y suele presentar un cierto grado de indefinición: si la hipótesis alternativa se formula simplemente como "*la hipótesis nula no es cierta*" el contraste es bilateral o a dos colas; por el contrario cuando se indica el sentido de la diferencia, el contraste es unilateral o a una sola cola. La decisión de rechazar o no la H_0 están al fin y al cabo basado en la elección de una muestra tomada al azar, y por tanto es posible cometer decisiones erróneas. Los errores que se pueden cometer se clasifican como sigue:

5.1.1. Error de tipo I

El error de tipo I es el error en que se incurre al rechazar H_0 cuando es cierta. La probabilidad de cometer este error es lo que se llama nivel de significación. Es una costumbre establecida denotarlo siempre con la letra α

$$\begin{aligned}\alpha &= \Pr\left(\textit{rechazar } H_0 \mid H_0 \textit{ es cierta}\right)\\ &= \Pr\left(\textit{aceptar } H_1 \mid H_0 \textit{ es cierta}\right)\end{aligned}$$

5.1.2. Error de tipo II

El error de tipo II es el error en que se incurre al no rechazar H_0 cuando es falsa. La probabilidad de cometer este error se denota con la letra β:

$$\begin{aligned}\beta &= \Pr\left(\textit{no rechazar } H_0 \mid H_0 \textit{ es falsa}\right)\\ &\neq \Pr\left(\textit{no rechazar } H_0 \mid H_1 \textit{ es cierta}\right)\end{aligned}$$

En ambos casos, se ha producido un juicio erróneo. Para que las reglas de decisión sean buenas, deben diseñarse de modo que minimicen los errores de la decisión; y no es una cuestión sencilla, porque para cualquier tamaño de la muestra, un intento de disminuir un tipo de error suele ir acompañado de un crecimiento del otro tipo. En la práctica, un tipo de error puede ser más grave que el otro, y debe alcanzarse

un compromiso que disminuya el error más grave. La única forma de disminuir ambos a la vez es aumentar el tamaño de la muestra, lo que no siempre es posible.

5.1.3. Tipos de contraste y error gamma fatal

La interpretación de la pruebas estadísticas está condicionada por la construcción de las hipótesis. Se pueden formular hipótesis alternativas en una o dos direcciones con relación a la hipótesis nula. Considérese el caso en donde se comparan dos tratamientos. Con un tipo de contraste unilateral se establecerían supuestos de superioridad o inferioridad de un tratamiento con respecto al otro. Por contra en una aproximación bilateral se establecerían diferencias entre ambos tratamientos, no importa en que sentido.

Considérese una aproximación bilateral en la comparación de dos tratamientos S y P:

- La H_0 establece que la diferencia es cero $H_0 : S-P=0$.
- La hipótesis alternativa $H_1 : S-P\neq 0$, la diferencia entre ambos tratamientos es diferente de cero.

La hipótesis alternativa se aceptará tanto si $A>B$ como si $B>A$. Supóngase que la realidad es $A>B$. La variabilidad muestral puede causar una mala jugada. Aún con el empleo de técnicas de muestreo adecuadas, es improbable pero posible que en la muestra resulte que $B\gg A$. La decisión será aceptar la hipótesis alternativa ¡pero precisamente en la dirección errónea, declarando que $B<A$! La probabilidad de cometer este error en la decisión es el riesgo gamma. Este es muy pequeño pero comparativamente mucho mas importante y dramático. Como siempre nada hará sospechar de la fiabilidad de los resultados y de la bondad de la muestra.

5.1.4. Niveles de significación

Se llama nivel de significación a la máxima probabilidad con la que se está dispuesto a correr el riesgo de cometer un error de tipo I al contrastar una hipótesis. Esta probabilidad, que a menudo se denota por α, suele especificar antes de tomar la muestra, de manera que los resultados obtenidos no influyan en su elección. En la práctica, es frecuente un nivel de significación de 0,05, 0,01 ó 0.001, si bien son posibles otros valores. Por ejemplo, si se escoge el nivel de significación 0,05 (ó 5 %) al diseñar una regla de decisión, entonces hay unas cinco oportunidades entre 100 de rechazar la hipótesis cuando debiera haberse aceptado. Es decir, se tiene un 95 % de confianza de que se ha adoptado la decisión correcta. En este caso se dice que la hipótesis ha sido rechazada al nivel de significación 0,05, que significa que tal hipótesis tiene una probabilidad 0,05 de ser falsa. El nivel de significación y el grado de confianza, están en relación inversa.

En un contraste de hipótesis (test de hipótesis, contraste de significación) se decide si cierta hipótesis H_0 puede ser rechazada o no a la vista de los datos suministrados por una muestra de la población. El procedimiento general consiste en:

(a) Planificar la H_0 y la H_1. La H_0 es el valor hipotético del parámetro que se compara con el resultado muestral resulta muy poco probable cuando la hipótesis es cierta. La H_1 será admitida cuando H_0 sea rechazada. Habitualmente H_1 es la negación de H_0, aunque esto no es necesariamente así;

(b) Especificar el nivel de significación que se va a utilizar. El intervalo de aceptación (o más exactamente de no rechazo de la H_0) se establece fijando el nivel de significación α, una cantidad suficientemente pequeña de modo que la probabilidad de que el estadístico del contraste tome un valor fuera del mismo –región crítica–

$$\textit{región crítica} \;\equiv\; C \;=\; \mathbb{R} \mid \left(T_i, T_s\right)$$

Por ejemplo, el nivel de significación del 5 % significa que se rechaza H_0 solamente si el resultado muestral es tan diferente del valor hipotético que una diferencia de esa magnitud o mayor, pudiera ocurrir aleatoriamente con una probabilidad de 0,05 o menos. El nivel de significación va a permitir separar sucesos que tienen mucha probabilidad de ocurrir de aquellos que la tienen baja: es una forma de marcar el error máximo que se está dispuesto a admitir;

(c) Definir un estadístico (*de prueba*, *de contraste*) T relacionado con la hipótesis que se desea contrastar. El estadístico de prueba puede ser el estadístico muestral (el estimador no segado del parámetro que

se prueba) o una versión transformada de ese estadístico muestral. Por ejemplo, para probar el valor hipotético de una media poblacional, se toma la media de una muestra aleatoria de esa distribución normal, entonces es común que se transforme la media en un valor z el cual, a su vez, sirve como estadístico de prueba;

(d) Establecer el valor o valores críticos del estadístico de prueba. Habiendo especificado la H_0, el nivel de significación y el estadístico de prueba que se van a utilizar, se procede a establecer el o los valores críticos del estadístico de prueba. Puede haber uno o más de esos valores, dependiendo de si se va a realizar una prueba de uno o dos extremos.

(e) Determinar el valor real del estadístico de prueba. Por ejemplo, al probar un valor hipotético de la media poblacional, se toma una muestra aleatoria y se determina el valor de la media muestral. Si el valor crítico que se establece es un valor de z, entonces se transforma la media muestral en un valor de z.

(f) Suponiendo que H_0 sea verdadera, calcular un intervalo (de aceptación de H_0,) $\left(T_i - T_s\right)$ de manera que al calcular sobre la muestra $T = T_{\exp}$ el criterio a seguir sea:

$$\begin{cases} Si\ T_{\exp} \in \left(T_i, T_s\right) \Rightarrow no\ se\ rechaza\ H_0\ \left(\neq se\ rechaza\ H_1\right) \\ Si\ T_{\exp} \notin \left(T_i, T_s\right) \Rightarrow se\ rechaza\ H_0\ y\ se\ acepta\ H_1 \end{cases}$$

(g) Toma de decisión, comparando el valor observado del estadístico muestral con el valor (o valores) críticos del estadístico de prueba y consecuentemente aceptar o rechazar H_0..

decisiones posibles	situaciones posibles: la hipótesis nula es verdadera	situaciones posibles: la hipótesis nula es falsa
aceptar la hipótesis nula	se acepta correctamente	error tipo II
rechazar la hipótesis nula	error tipo I	se rechaza correctamente

Tabla I. Consecuencias de las decisiones en pruebas de hipótesis

5.2. Pruebas de uno y dos extremos (unilaterales y bilaterales, de una y dos colas)

Cuando se estudian ambos valores estadísticos es decir, ambos lados de la media se denomina prueba de uno y dos extremos o contraste de una y dos colas. No obstante, se estará interesado con frecuencia tan sólo en valores extremos a un lado de la media (o sea, en uno de los extremos de la distribución), como sucede cuando se contrasta la hipótesis de que un proceso es mejor que otro (que es lo mismo que contrastar si un proceso es mejor o peor que otro). Tales contrastes se llaman unilaterales, de un extremo o de una cola. En esta situación, la región crítica es una región situada a un lado de la distribución, con área igual al nivel de significación.

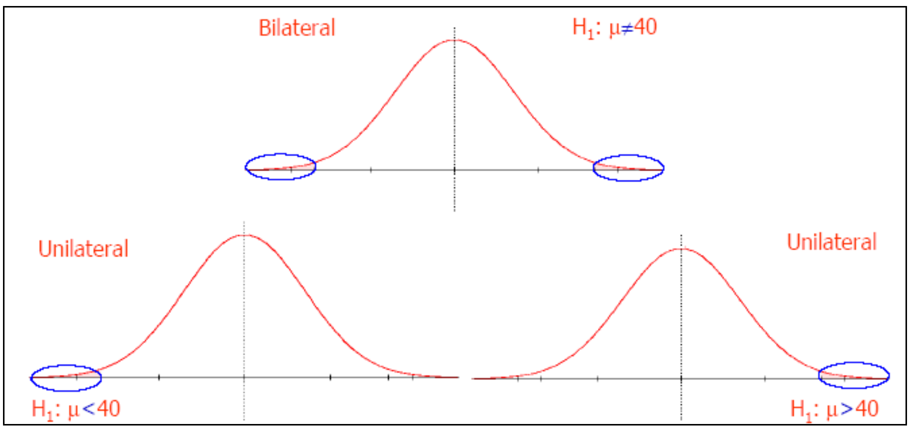

Figura 4. Contrastes unilateral y bilateral. La posición de la región crítica depende de la hipótesis alternativa.

	nivel de significación α				
valores críticos de z	**0,10**	**0,05**	**0,01**	**0,005**	**0,02**
tests unilaterales	–1,28 o 1,28	–1,645 o 1,645	–2,33 o 2,33	–2,58 o 2,58	–2,88 o 2,88
tests bilaterales	– 1,645 y 1,645	–1,96 y 1,96	–2,58 y 2,58	–2,81 y 2,81	–3,08 y 3,08

Tabla II. valores críticos de z para contraste de unos o dos extremos en varios niveles de significación

5.3. Curva característica operativa y curva de potencia

Se ha visto como limitar el error de tipo I eligiendo adecuadamente el nivel de significación. Es posible evitar el riesgo de cometer error de tipo II simplemente no aceptado nunca hipótesis, pero en muchas aplicaciones prácticas esto es inviable. En tales casos se suele recurrir a curvas de operación características (OC), gráficos que muestran las probabilidades de error de tipo II bajo diversas hipótesis. Proporcionan indicadores de hasta que punto un test permitirá evitar un error de tipo II; es decir, indicará la potencia de un test para prevenir decisiones erróneas. Son útiles en el diseño de experimentos porque sugieren entre otras cosas al tamaño de muestra a utilizar.

5.4. Grados de libertad

Para el cálculo de un estadístico, es necesario emplear tanto observaciones de muestra como propiedades de ciertos parámetros de la población. Si estos parámetros son desconocidos, hay que estimarlos a partir de la muestra el número de grados de libertad de un estadístico, generalmente denotado por ν, y definido como el número N de observaciones independientes en la muestra (el tamaño de la muestra) menos el número k de parámetros de la población, estimado a partir de observaciones muéstrales. Simbólicamente:

$$\nu = N - k \, .$$

5.5. Observaciones

1. Los errores de tipo I y II no están relacionados más que del siguiente modo: cuando α decrece β crece. Por tanto no es posible encontrar tests que hagan tan pequeños como se quiera ambos errores simultáneamente. De modo que siempre es necesario privilegiar a una de las hipótesis, que no será rechazada, a menos que su falsedad se haga muy evidente. En los contrastes, la hipótesis privilegiada es H_0 que sólo será rechazada cuando la evidencia de su falsedad supere el umbral $100(1-\alpha\,\%)$.

2. Al tomar α muy pequeño se tendrá que β puede aproximarse a 1. Lo ideal a la hora de definir un test es encontrar un compromiso satisfactorio entre α y β (aunque siempre a favor de H_0)

Se denomina *potencia o poder estadístico de un contraste* a la cantidad $1-\beta$, es decir:

$$potencia\ del\ contraste \equiv 1-\beta = \Pr\left(rechazar\ H_0 \mid H_0\ es\ falsa\right)$$

	no rechazar H_0	**rechazar H_0**
H_0 es cierta	correcto probabilidad $1-\alpha$	error tipo I probabilidad α
H_0 es falsa	error tipo II probabilidad β	correcto probabilidad $1-\beta$

Tabla III.

3. En el momento de elegir una hipótesis privilegiada se puede, en principio, dudar entre si elegir una dada o bien su contraria. Los siguientes criterios son a tener en cuenta en estos casos:

- Simplicidad científica: a la hora de elegir entre dos hipótesis científicamente razonables, se tomará como H_0 la más simple;

- Las consecuencias de equivocarse: por ejemplo, al juzgar el efecto que puede causar cierto tratamiento médico que está en fase de experimentación, en principio se ha de tomar como H_0 aquella cuyas consecuencias de no rechazarla siendo falsa son menos graves, y como H_1 aquella en la que el aceptarla siendo falsa trae peores consecuencias. Es decir:

$$\begin{cases} H_0 \ : \ el \ paciente \ empeora \ o \ queda \ igual \ ante \ el \ tratamiento \\ H_1 \ : \ el \ paciente \ mejora \ con \ el \ tratamiento \end{cases}$$

Ejemplo 2

Otro ejemplo claro: cuando se acaban de instalar un nuevo ascensor en el edificio y se quiere saber si caerá o no al vacío cuando estén personas dentro. Una persona prudente es la que espera a que un número suficiente de vecinos suyos hayan usado el ascensor (muestra aleatoria) y realiza un test del tipo:

$$\begin{cases} H_0 \ : \ el \ ascensor \ se \ caerá \\ H_1 \ : \ el \ ascensor \ no \ se \ caerá \end{cases}$$

y sólo aceptará la H_1 para $\alpha \approx 0$ aunque para ello tenga que ocurrir que $\beta \approx 1$, ya que las consecuencias del error de tipo I (ir al hospital) son mucho más graves que las del error del tipo II (subir a pie varios pisos). Es decir a la hora de decidirse por una de las dos hipótesis no basta con elegir la más probable (nadie diría "*voy a tomar el ascensor pues la probabilidad de que no se caiga es del 60 %*"). Hay que elegir siempre la hipótesis H_0 a menos que la evidencia a favor de H_1 sea muy significativa.

Ejemplo 1. (continuación)

Volviendo al ejemplo de la estatura de los habitantes de un pueblo, un estadístico de contraste adecuado es $\bar{X}$. Si la hipótesis H_0 fuese cierta se tendría que

$$\bar{X} \ \rightarrow \ N\left(\mu \ , \frac{\sigma^2}{n}\right)$$

(Suponiendo, claro está, que la distribución de las alturas de los españoles siga una distribución normal de parámetros conocidos, por ejemplo, $N\left(\mu = 1,74 \ ; \ \sigma^2 = 10^{-2}\right)$. Si μ_0 es el verdadero valor de la media en el pueblo en estudio, como la varianza de $\bar{X}$ es pequeña para grandes valores de n, lo lógico es pensar que si el valor obtenido con la muestra $\bar{X} = \bar{x}$ está muy alejado de $\mu = 1,74$ (región crítica), entonces:

- o bien la muestra es muy extraña, si H_0 es cierta (probabilidad α); o bien
- la hipótesis H_0 no es cierta.

Concretamente en el primer caso, donde la muestra es $\{1,50, 1,52, 1,48, 1,55, 1,60, 1,49, 1,55, 1,63\}$, el contraste de hipótesis conveniente es:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$

Aquí H_1 no es estrictamente la negación de H_0. Esto dará lugar a un *contraste unilateral*, que es aquel en los que la región crítica está formada por un sólo intervalo:

$$intervalo \ de \ no \ rechazo \ de \ H_0 \ \equiv \ \left(T_i \ , +\infty\right)$$
$$región \ crítica \ \equiv \ \left(-\infty \ , T_i\right]$$

En el segundo caso, donde la muestra es $\{1,65, 1,80, 1,73, 1,52, 1,75, 1,65, 1,75, 1,78\}$, el contraste de hipótesis que debería realizarse es:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

Aquí sí se puede decir que H_1 es la negación de H_0: se trata de un *contraste bilateral***,** aquel en el que la región crítica está formada por dos intervalos separados:

$$\textit{intervalo donde no se rechaza la } H_0 \equiv \left(T_i \, , T_s\right)$$

$$\textit{región crítica} \equiv \left(-\infty \, , T_i\right] \cup \left[T_s \, , +\infty\right)$$

Los últimos conceptos que van a introducirse son:

- *Hipótesis simple*: aquella en la que se especifica un único valor del parámetro. Es el caso de las hipótesis nulas en los dos últimos contrastes mencionados;
- *Hipótesis compuesta*: aquella en la que se especifica más de un posible valor del parámetro. Por ejemplo, son compuestas las hipótesis alternativas de esos mismos contrastes.

5.6. El concepto de valor p o grado de significación

Supóngase que, para una muestra dada, un test de hipótesis rechaza la hipótesis nula H_0 a un cierto nivel de significación α. Si se va bajando poco a poco el nivel de significación (manteniendo los mismos datos) el test se va haciendo más y más conservador a favor de H_0, de manera que H_0 se va rechazando "cada vez por menor margen" hasta que llega un momento en que la hipótesis nula se acepta a partir de un determinado valor de $\alpha = p$, y también para todos los valores de α menores que él.

Para una muestra dada, valor p o grado de significación, de un test, es aquel que viene dado por los datos del experimento y es la probabilidad de encontrar una diferencia igual o superior a la hallada cuando la hipótesis nula es cierta. El valor p se interpreta como una medida de la evidencia estadística que los datos aportan a favor de la hipótesis alternativa H_1 (o en contra de H_0): Cuando el valor p es "muy pequeño" (por ejemplo, $\leq 0,01$) se considera que hay una fuerte evidencia a favor de H_1 ya que ha sido necesario bajar mucho el nivel de significación para poder aceptar H_0. En otras palabras: el valor p indica "el punto de división" entre el rechazo y la aceptación: Si se aplica un test con un nivel más alto que el valor p se rechazaría H_0, si el nivel de significación es más pequeño, se acepta. Los programas estadísticos informáticos habituales suelen proporcionar el valor p de los diferentes tests. Esta es una información muy completa ya que el usuario puede conocer el resultado (aceptación o rechazo de H_0) para todos los posibles valores de α.

5.7. Errores comunes en la interpretación del nivel de significación y el valor p

La investigación sobre la comprensión de los métodos de inferencia muestra la existencia de concepciones erróneas ampliamente extendidas, tanto entre los estudiantes universitarios, como entre los científicos que usan la inferencia estadística en su trabajo diario. Estas concepciones erróneas se refieren principalmente al nivel de significación α, que se define como la probabilidad de rechazar la hipótesis nula en caso de que sea cierta.

5.7.1. Inferencia y probabilidad condicional

La interpretación errónea más extendida de este concepto consiste en intercambiar los dos términos de la probabilidad condicional, es decir, en interpretar el nivel de significación como la probabilidad de que la hipótesis nula sea cierta si se ha tomado la decisión de rechazarla. Por ejemplo:

- Birnbaum (1982), informó que sus estudiantes encontraban razonable la siguiente definición: "*Un nivel de significación del 5 % indica que, en promedio, 5 de cada 100 veces que se rechace la hipótesis nula se estará equivocado"* (1);
- Falk (1986) comprobó que la mayoría de sus estudiantes creían que α era la probabilidad de equivocarse al rechazar la hipótesis nula (2);

- Vallecillos (1994) planteó las siguientes cuestiones a una muestra de 436 estudiantes universitarios de diferentes especialidades que habían estudiado el tema (3-5):
  1. Un nivel de significación del 5 % significa que, en promedio, 5 de cada 100 veces que se rechace la hipótesis nula se estará equivocado (verdadero o falso?: justifíquelo).
  2. Un nivel de significación del 5 % significa que, en promedio, 5 de cada 100 veces que la hipótesis nula es cierta será rechazada (verdadero o falso?: justifíquelo).

  En la cuestión 2 se presenta una(correcta) interpretación frecuencial del nivel de significación, mientras que en la cuestión 1 se han intercambiado (incorrectamente) los dos sucesos que definen la probabilidad condicional. Sin embargo, sólo el 32 % de los estudiantes dio una respuesta correcta a la primera cuestión y el 54 % dio una respuesta correcta a la segunda. De 135 estudiantes que justificaron su respuesta, el 41 % dio un argumento correcto en los dos ítems. Un error prevalente en todos los grupos de estudiantes fue el intercambio de los términos de la probabilidad condicional, juzgando por tanto correcto la primera cuestión y falsa la segunda. Entrevistas a un grupo reducido de estudiantes mostró que esta creencia aparecía en algunos estudiantes que eran capaces de discriminar entre una probabilidad condicional y su inversa (6). Otros estudiantes no distinguían las dos probabilidades condicionales, es decir, consideraban que ambos ítems eran correctos.

Ejemplo 3

Que las probabilidades condicionales con términos intercambiados no coinciden, en general, se ilustra en la tabla IV que se refiere a la elección de estadística como tema optativo en una escuela. La probabilidad de que una chica tomada al azar estudie estadística y la probabilidad de que un estudiante de estadística sea una chica son diferentes:

$$\Pr\left(\textit{estudie estadística} \mid \textit{chica}\right) = \frac{3}{4}$$

$$\Pr\left(\textit{chica} \mid \textit{estudie estadística}\right) = \frac{3}{8}$$

	chicas	chicos	total
estadística	300	500	800
no estadística	100	100	200
total	400	600	1000

Tabla IV. Número de chicos y chicas en un curso de estadística

Es importante resaltar que, incluso cuando se fija el nivel de significación α, es decir, la probabilidad de rechazar la H_0 (supuesto que es cierta) y se pueda calcular la probabilidad de obtener un valor del estadístico de contraste menor que un valor particular (supuesta H_0 cierta), la probabilidad de que H_0 sea cierta una vez sea rechazada y la probabilidad de que H_0 sea cierta una vez que se ha obtenido el valor del estadístico de contraste no pueden conocerse.

La probabilidad a posteriori de H_0 dado un resultado significativo depende de la probabilidad a priori de H_0, así como de las probabilidades de obtener un resultado significativo, dadas las H_0 y H_1. Desafortunadamente estas probabilidades no pueden determinarse. Más aún, una hipótesis es o cierta o falsa y, por tanto, no tiene mucho sentido calcular su probabilidad en un paradigma inferencial clásico (donde se da una interpretación frecuencial a las probabilidades objetivas). Sólo en la inferencia bayesiana pueden calcularse las probabilidades a posteriori de la hipótesis, aunque son probabilidades subjetivas, Lo más que se puede hacer, y es usando inferencia bayesiana, es revisar el grado de creencia personal en la hipótesis en vista de los resultados.

5.7.2. Otras interpretaciones erróneas del nivel de significación y el valor *p*

Hay quien piensa que el valor *p* es la probabilidad de que el resultado se deba al azar. Puede verse claramente que esta concepción es errónea en el hecho de que incluso si H_0 es cierta (por ejemplo, si no hubiera diferencias de rendimiento en el ejemplo 1) un resultado significativo puede ser debido a otros

factores, como, por ejemplo, que los estudiantes del grupo experimental trabajasen más que sus compañeros al prepararse para la evaluación. De aquí la importancia del control experimental para intentar asegurar que todas las condiciones (excepto el tipo de enseñanza) se mantienen constantes en los dos grupos. El valor *p* es la probabilidad de obtener el resultado particular u otro más extremo cuando la hipótesis nula es cierta *y* no hay otros factores posibles que influencien el resultado. Lo que se rechaza en un contraste de hipótesis es H_0 y, por tanto, no se puede inferir la existencia de una causa particular en un experimento a partir de un resultado significativo.

Otro error común es la creencia en la conservación del valor del nivel de significación cuando se realizan contrastes consecutivos en el mismo conjunto de datos, lo que produce el problema de las comparaciones múltiples. A veces se aplica un gran número de pruebas de significación a un mismo conjunto de datos. El significado del nivel de significación es que, si se llevan a cabo 100 comparaciones sobre el mismo conjunto de datos y se usan en todas ellos el nivel de significación 0,05, habrá que esperar que 5 de las 100 pruebas sean significativas por puro azar, incluso cuando la hipótesis nula sea cierta, con la consiguiente dificultad para interpretar los resultados (Moses, 1992).

El habitual uso de los niveles de significación 0,05 y 0,01 es cuestión de convenio y no se justifica por la teoría matemática. Si se considera el contraste de hipótesis como proceso de decisión (la visión de Neyman y Pearson), debe especificarse el nivel de significación antes de llevar a cabo el experimento y esta elección determina el tamaño de las regiones críticas y de aceptación que llevan a la decisión de rechazar o no la hipótesis nula. Neyman y Pearson dieron una interpretación frecuencial a esta probabilidad: Si la H_0 es cierta y se repite el experimento muchas veces con probabilidad de error tipo I igual a 0,05 se rechazará la hipótesis nula el 5 % de las veces que sea cierta. Inicialmente Fisher (1935) sugirió seleccionar un nivel de significación del 5 %, como convenio para reconocer los resultados significativos en los experimentos. Posteriormente, el mismo Fisher (1956) consideró que cada investigador debe seleccionar el nivel de significación de acuerdo a las circunstancias, ya que "de hecho ningún investigador mantiene un nivel de significación fijo con el cual rechaza las hipótesis año tras año y en todas las circunstancias", sugiriendo que se publicara el valor p exacto obtenido en cada experimento particular, lo que de hecho, implica establecer el nivel de significación después de llevar a cabo el experimento (7). A pesar de estas recomendaciones, la literatura de investigación muestra que los niveles arbitrarios de 0,05, 0,01, 0,001 se usan casi en forma universal para todo tipo de problemas. A pesar de que esta práctica introduce sesgo de publicación (8). A veces, si la potencia del contraste es baja y el error tipo II es importante, sería preferible una probabilidad mayor de error tipo I.

A la interpretación incorrecta del nivel de significación se une habitualmente una interpretación incorrecta de los resultados significativos, punto también de desacuerdos entre Fisher y Neyman y Pearson. Para Fisher un resultado significativo implica que los datos proporcionan evidencia en contra de H_0, mientras que para Neyman y Pearson solo establece la frecuencia relativa de veces que se rechazaría H_0 cierta a la larga (error tipo I). Por otro lado, hay que diferenciar entre significación estadística y significación práctica. En el ejemplo 1 se obtuvo una diferencia media significativa en puntuaciones entre los dos grupos de 13,32. Sin embargo, se podría haber obtenido una significación estadística mayor con un efecto experimental menor y una muestra de tamaño mayor. La significación práctica implica significación estadística más un efecto experimental suficientemente elevado.

5.8. Los valores p *y los intervalos de confianza: ¿en qué confiar?*

Dado un contraste bilateral

$$\begin{cases} H_0 \ : \ \theta = \theta_0 \\ H_1 \ : \ \theta \neq \theta_0 \end{cases}$$

con nivel de significación α, se rechaza la H_0 si θ_0 no pertenece al intervalo de confianza para θ con nivel de confianza $1-\alpha$.

La mayoría de los estadísticos desaconsejan el uso de las pruebas de hipótesis estadísticas debido a las graves deficiencias de estas pruebas y a su dudosa utilidad en comparación con otros métodos de análisis inferencial. Los primeros argumentos en contra del uso de pruebas de hipótesis aparecieron durante la primera mitad del siglo pasado con carácter esporádico, pero cuando en 1986 *British Medical Journal* dio a conocer su postura al respecto, el debate en torno a ellas cobró un ímpetu que ha quedado evidenciado en los centenares de artículos posteriores que contrastan las ventajas de los intervalos de confianza con las

carencias del valor p (9). Lo cierto es que estas últimas, raras veces abordadas en la sala de clase, son considerables desde el punto de vista de los fines que persigue un investigador. Hoy por hoy se reconoce que los intervalos de confianza aventajan a las pruebas de hipótesis como instrumento analítico para muchos tipos de investigación, entre ellos los estudios observacionales y experimentales relacionados con las ciencias médicas y sociales, con el resultado de que la mayoría de las revistas biomédicas alientan a sus autores a proporcionar intervalos de confianza en lugar de valores p. Fáciles de calcular con los paquetes estadísticos modernos, los valores p ejercen un poderoso atractivo sobre el investigador por la exigua reflexión que exigen y la falsa sensación de seguridad que confieren. Un solo número encierra la clave que determina si los resultados de un estudio han de sumarse a las pruebas a favor o en contra de una hipótesis, y el investigador que obtiene resultados significativos suele sentirse satisfecho de haber logrado su meta, sin darse cuenta de que no ha conseguido mejorar en modo alguno su comprensión del fenómeno que estudia. Para entender a fondo esta afirmación, conviene examinar qué es un valor p.

5.9. Potencia o poder estadístico de un estudio

5.9.1. Factores que influyen en la potencia o poder estadístico de un estudio

La potencia o *poder estadístico* de un estudio depende de diferentes factores, como:

- *El tamaño del efecto a detectar*, es decir, la magnitud mínima de la diferencia o asociación entre los grupos que se considera clínicamente relevante. Cuanto mayor sea el tamaño del efecto que se desea detectar, mayor será la probabilidad de obtener hallazgos significativos y, por lo tanto, mayor será el poder estadístico;
- La *variabilidad* de la respuesta estudiada: cuanto mayor sea la variabilidad en la respuesta, más difícil será detectar diferencias entre los grupos que se comparan y menor será el poder estadístico de la investigación. De ahí que sea recomendable estudiar grupos lo más homogéneos posibles;
- El *tamaño de la muestra* a estudiar: cuanto mayor sea el tamaño muestral, mayor será la potencia estadística de un estudio. Es por ello que en los estudios con muestras muy grandes se detectan como significativas diferencias poco relevantes, y en los estudios con muestras menores es más fácil obtener resultados falsamente negativos;
- El *nivel de significación estadística*. Si se disminuye el valor de α también se disminuye el poder de la prueba. Es decir, si se disminuye la probabilidad de cometer un error de tipo I aumenta simultáneamente la probabilidad de un error de tipo II, por lo que se trata de encontrar un punto de "equilibrio" entre ambas. Habitualmente se trabaja con un nivel de significación del 95 % $(\alpha = 0,05)$, por lo que el equilibrio hay que en encontrarlo finalmente entre el tamaño de la muestra que es posible estudiar y la potencia requerida para el estudio.

Los cuatro factores anteriores, junto con el poder estadístico, forman un sistema cerrado. De este modo, una vez fijados tres de ellos, el cuarto queda completamente determinado.

5.9.2. Cálculo del poder estadístico de un estudio

A la hora de diseñar una investigación, es importante determinar si dicho estudio alcanzará una precisión suficiente. Generalmente, se suele trabajar con una potencia en torno al 80 % o al 90 %. Sin embargo, las condiciones en las que se lleva a cabo con frecuencia una investigación son diferentes de las que se habían previsto en un principio. En consecuencia, y a la vista de hallazgos no significativos, es recomendable evaluar de nuevo a posteriori su potencia con el fin de discernir si el estudio carece del poder necesario para detectar una diferencia relevante o bien si realmente puede no existir tal diferencia.

En la tabla VI se muestran las fórmulas necesarias para el cálculo del poder estadístico en función de la naturaleza de la investigación. Estas fórmulas permiten obtener un valor $z_{1-\beta}$ a partir del cual se puede determinar el poder asociado recurriendo a las tablas de la distribución normal. En la tabla VII se muestra la correspondencia entre algunos valores de $z_{1-\beta}$ y el poder estadístico asociado. Sin embargo, y aunque dichas fórmulas permitirían analizar la potencia estadística en diferentes tipos de diseño, puede resultar más sencillo disponer de algún *software* específico con el que poder realizar dichos cálculos

Ejemplo 4

Supóngase que se quiere llevar a cabo un ensayo clínico para comparar la efectividad de un nuevo fármaco con la de otro estándar en el tratamiento de una determinada enfermedad. Al inicio del estudio, se sabe que

la eficacia del tratamiento habitual está en torno al 40 %, y se espera que con el nuevo fármaco la eficacia aumente al menos en un 15 %. El estudio se diseñó para que tuviese una potencia del 80 %, asumiendo una seguridad del 95 %. Esto implica que son necesarios 173 pacientes en cada uno de los grupos para llevar a cabo la investigación. Tras finalizar el estudio, sólo fue posible tratar con cada uno de los fármacos a 130 pacientes en cada grupo en lugar de los 173 pacientes estimados inicialmente. Al realizar el análisis estadístico, se objetivó que no hay diferencias significativas en la efectividad de ambos tratamientos. A partir de las fórmulas de la Tabla VI, se puede calcular la potencia final del estudio. Aplicando la fórmula para el cálculo del poder estadístico de comparación de dos proporciones ante un planteamiento unilateral se obtiene:

$$\left.\begin{array}{l} \left.\begin{array}{l} p_1 = 0,40 \\ p_1 = 0,55 \end{array}\right\} \Rightarrow\ p = 0,475 \\ n = 130 \\ \alpha = 0,05 \Rightarrow z_{1-\alpha} = 1,645 \end{array}\right\} \Rightarrow z_{1-\beta} = \frac{\left|p_1 - p_2\right|\sqrt{n} - z_{1-\alpha}\sqrt{2\,p\left(1-p\right)}}{\sqrt{p_1\left(1-p_1\right) + p_2\left(1-p_2\right)}} =$$

$$= \frac{\left|0,44-0,55\right|\sqrt{130} - 1,654\sqrt{2\times 0,475\times\left(1-0,475\right)}}{\sqrt{0,4\times\left(1-0,4\right) + 0,55\times\left(1-0,55\right)}} =$$

$$= 0,467$$

A partir de la tabla VII, se puede determinar que un valor de $z_{1-\beta} = 0,467$ corresponde a una potencia en torno al 65–70 %. Utilizando las tablas de la distribución normal, se sabe que la potencia es del 68 %, es decir, el estudio tendría un 68 % de posibilidades de detectar una mejora en la eficacia del tratamiento del 15 %.

Utilizando la fórmula anterior, podría obtenerse un gráfico como en el que se muestra en la figura 5, en la que, para este ejemplo, se estima la potencia estadística del estudio en función del tamaño de la muestra estudiada y la magnitud del efecto a detectar. Así, puede concluirse que de haber estudiado 130 pacientes por grupo, se obtiene una potencia de sólo el 36,6 % para detectar una diferencia mínima del 10 %, una potencia del 68 % para detectar una diferencia del 15 % y de un 90,2 % para una diferencia del 20 %. Este tipo de gráficos resulta muy útil tanto en la fase de diseño de un estudio como a la hora de valorar a posteriori el poder de una investigación.

Ejemplo 5

Supóngase que se quiere llevar a cabo un estudio de casos y controles para estudiar la posible asociación entre la presencia de cardiopatía isquémica y el hábito de fumar. De acuerdo con estudios previos, se cree que la incidencia de cardiopatía puede ser hasta dos veces más alta entre los fumadores, y se asume que la frecuencia de exposición entre los controles será de un 40 %. Debido a ciertas limitaciones, sólo es posible para el investigador incluir en el estudio a 100 pacientes con cardiopatía isquémica (casos). Utilizando las fórmulas de la tabla VI, con un planteamiento bilateral y una seguridad del 95 %:

$$\left.\begin{array}{l} OR = 2 \\ p_2 = 0,4 \end{array}\right\} \Rightarrow\ p_1 = \frac{OR\ p_2}{\left(1-p_2\right)+OR\,p_2} = \frac{2\times 0,4}{\left(1-0,4\right)+2\times 0,4} = 0,5714$$

$$p = \frac{p_1 + p_2}{2} = 0,486$$

$$n = 100$$

$$\alpha = 0,05 \Rightarrow z_{1-\alpha/2} = 1,96$$

$$c = 1 \Rightarrow$$

$$\Rightarrow z_{1-\beta} = \frac{\left|p_1 - p_2\right|\sqrt{c\,n} - z_{1-\alpha}\sqrt{(c+1)\,p\,(1-p)}}{\sqrt{c\,p_1\left(1-p_1\right) + p_2\left(1-p_2\right)}} =$$

$$= \frac{\left|0{,}5174 - 0{,}40\right|\sqrt{100} - 1{,}96\sqrt{2\times 0{,}486\times\left(1-0{,}486\right)}}{\sqrt{0{,}5714\times\left(1-0{,}5714\right) + 0{,}40\times\left(1-0{,}40\right)}} =$$

$$= 0{,}472$$

En las tablas de la distribución normal se obtiene para un valor $z_{1-\beta} = 0{,}472$ una potencia del 68,17 %.

Con el fin de mejorar la potencia del estudio, los investigadores se plantean reclutar un mayor número de controles que de casos. En la figura 6 se muestra para el ejemplo anterior el poder de la investigación en función del número de casos y controles estudiados. Como se puede observar, la ganancia en la potencia disminuye rápidamente, y es prácticamente nula cuando la relación entre el número de controles y casos es 4:1. Esto se verifica en cualquier estudio de casos y controles. En particular, para el ejemplo previo, si se estudiasen 100 casos y 200 controles se alcanzaría una potencia del 80,28%. Si se incluyesen 100 casos y 300 controles, la potencia sería de un 84,69 %. Con 400 controles la potencia aumentaría sólo a un 86,89 % y con 500 a un 88,19 %. Con lo cual claramente es ineficiente el incluir más de 4 controles por caso ya que no se lograría un incremento relevante de la potencia estadística.

El análisis adecuado de la potencia estadística de una investigación, que es en definitiva la capacidad que tiene el estudio para encontrar diferencias si es que realmente las hay, es un paso fundamental tanto en la fase de diseño como en la interpretación y discusión de sus resultados. A la hora del diseño, por tanto, debe establecerse la magnitud mínima de la diferencia o asociación que se considere de relevancia clínica, así como la potencia estadística que se desea para el estudio y, de acuerdo con ello, calcular el tamaño de la muestra necesaria. Tras realizar el análisis estadístico, cuando se dice que no existe evidencia de que A se asocie con B o sea diferente de B, deberá cuestionarse antes de nada si la ausencia de significación estadística indica realmente que no existe una diferencia o asociación clínicamente relevante, o simplemente que no se dispone de suficiente número de pacientes para obtener hallazgos significativos. Tanto si los hallazgos son estadísticamente significativos como si no lo son, la estimación de intervalos de confianza pueden también facilitar la interpretación de los resultados en términos de magnitud y relevancia clínica, proporcionando una idea de la precisión con la que se ha efectuado al estimación, de la magnitud y de la dirección del efecto. De este modo, los intervalos de confianza permiten tener una idea acerca de la potencia estadística de un estudio y, por tanto, de la credibilidad de la ausencia de hallazgos significativos.

resultado de la prueba: asociación o diferencia		significativa	no significativa
realidad: asociación o diferencia:	existe	no error: $1-\beta$	error de tipo II: β
	no existe	error de tipo I: α	no error: $1-\alpha$

Tabla V. Posibles conclusiones tras una prueba estadística de contraste de hipótesis. α es la probabilidad de cometer un error de tipo I; β es la probabilidad de cometer un error de tipo II.

	test unilateral	test bilateral
comparación de dos proporciones	$z_{1-\beta} = \dfrac{\lvert p_1 - p_2 \rvert \sqrt{n} - z_{1-\alpha}\sqrt{2\,p(1-p)}}{\sqrt{p_1(1-p_1) + p_2(1-p_2)}}$	$z_{1-\beta} = \dfrac{\lvert p_1 - p_2 \rvert \sqrt{n} - z_{1-\alpha/2}\sqrt{2\,p(1-p)}}{\sqrt{p_1(1-p_1) + p_2(1-p_2)}}$
comparación de dos media	$z_{1-\beta} = \sqrt{\dfrac{n}{2}}\,\dfrac{d}{s} - z_{1-\alpha}$	$z_{1-\beta} = \sqrt{\dfrac{n}{2}}\,\dfrac{d}{s} - z_{1-\alpha/2}$
estimación de un OR en estudios de casos y controles	$p_1 = \dfrac{OR\ p_2}{(1-p_2) + OR\ p_2}$ $z_{1-\beta} = \dfrac{\lvert p_1 - p_2 \rvert \sqrt{n\,c} - z_{1-\alpha}\sqrt{(c+1)\,p(1-p)}}{\sqrt{c\,p_1(1-p_1) + p_2(1-p_2)}}$ $m = c \times n$	$p_1 = \dfrac{OR\ p_2}{(1-p_2) + OR\ p_2}$ $z_{1-\beta} = \dfrac{\lvert p_1 - p_2 \rvert \sqrt{n\,c} - z_{1-\alpha/2}\sqrt{(c+1)\,p(1-p)}}{\sqrt{c\,p_1(1-p_1) + p_2(1-p_2)}}$ $m = c \times n$
estimación de un RR	$p_1 = RR \times p_2$ $z_{1-\beta} = \dfrac{\lvert p_1 - p_2 \rvert \sqrt{n} - z_{1-\alpha}\sqrt{2\,p(1-p)}}{\sqrt{p_1(1-p_1) + p_2(1-p_2)}}$	$p_1 = RR \times p_2$ $z_{1-\beta} = \dfrac{\lvert p_1 - p_2 \rvert \sqrt{n} - z_{1-\alpha/2}\sqrt{2\,p(1-p)}}{\sqrt{p_1(1-p_1) + p_2(1-p_2)}}$
estimación de un coeficiente de correlación lineal	$z_{1-\beta} = \sqrt{n-3}\,\dfrac{1}{2}\ln\left(\dfrac{1+r}{1-r}\right) - z_{1-\alpha}$	$z_{1-\beta} = \sqrt{n-3}\,\dfrac{1}{2}\ln\left(\dfrac{1+r}{1-r}\right) - z_{1-\alpha/2}$

n Tamaño muestral. En un estudio de casos y controles, n es el número de casos

p_1 En un estudio transversal o de cohortes, proporción de expuestos que desarrollan la enfermedad. En un estudio de casos y controles, proporción de casos expuestos

p_2 En un estudio transversal o de cohortes, proporción de no expuestos que desarrollan la enfermedad. En un estudio de casos y controles, proporción de controles expuestos

$$p = \frac{p_1 + p_2}{2}$$

d Valor mínimo de la diferencia a detectar entre dos medias

s^2 Varianza en el grupo control o de referencia

c Número de controles por caso

m En un estudio de casos y controles, número de controles

OR Valor aproximado del *odds ratio* a detectar

RR Valor aproximado del riesgo relativo a detectar

r Magnitud del coeficiente de correlación a detectar

Tabla VI. Fórmulas para el cálculo del poder estadístico para diferentes tipos de diseño.

seguridad	α	test unilateral $z_{1-\alpha}$	test bilateral $z_{1-\alpha/2}$
80 %	0,200	0,842	1,282
85 %	0,150	1,036	1,440
90 %	0,100	1,282	1,645
95 %	0,050	1,645	1,960
97,5 %	0,025	1,960	2,240
99 %	0,010	2,326	2,576
poder estadístico	$1-\beta$	β	$z_{1-\beta}$
99 %	0,99	0,01	2,326
95 %	0,95	0,05	1,645
90 %	0,90	0,10	1,282
85 %	0,85	0,15	1,036
80 %	0,80	0,20	0,842
75 %	0,75	0,25	0,674
70 %	0,70	0,30	0,524
65 %	0,65	0,35	0,385
60 %	0,60	0,40	0,253
55 %	0,55	0,45	0,126
50 %	0,50	0,50	0,000

Tabla VII. Valores de $z_{1-\alpha}$, $z_{1-\alpha/2}$ y $z_{1-\beta}$ y más frecuentemente utilizados.

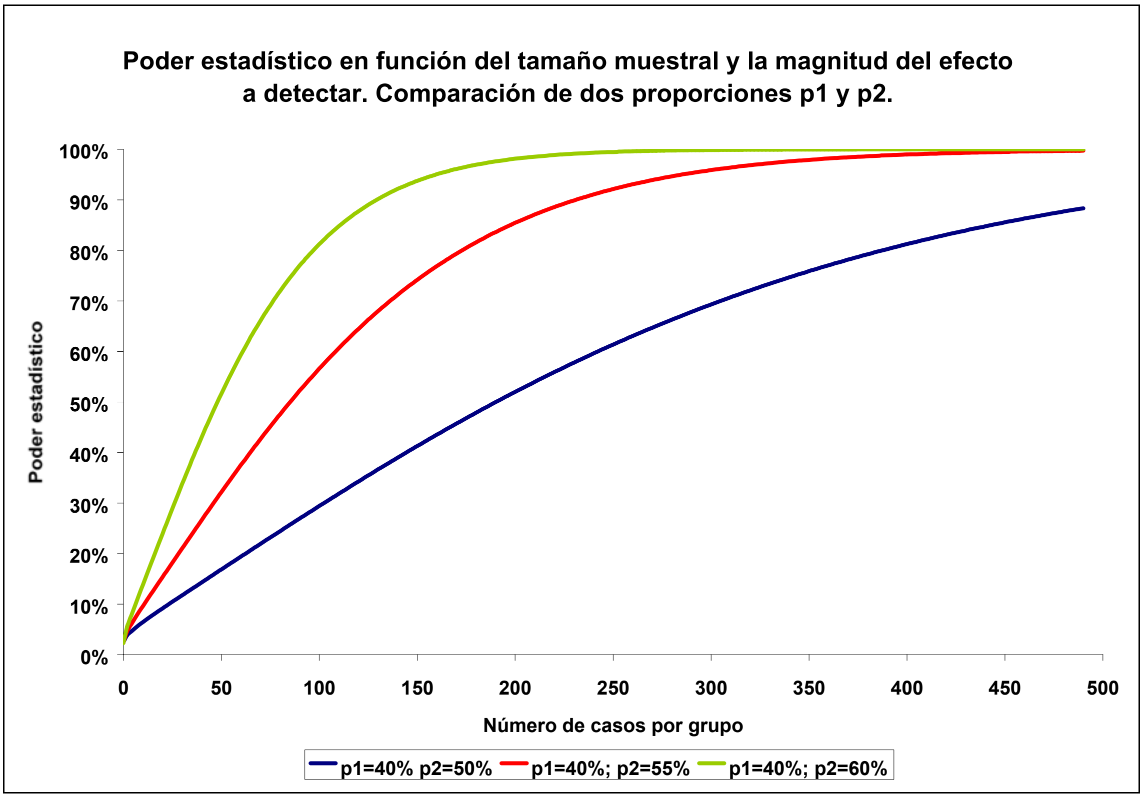

Figura 5.

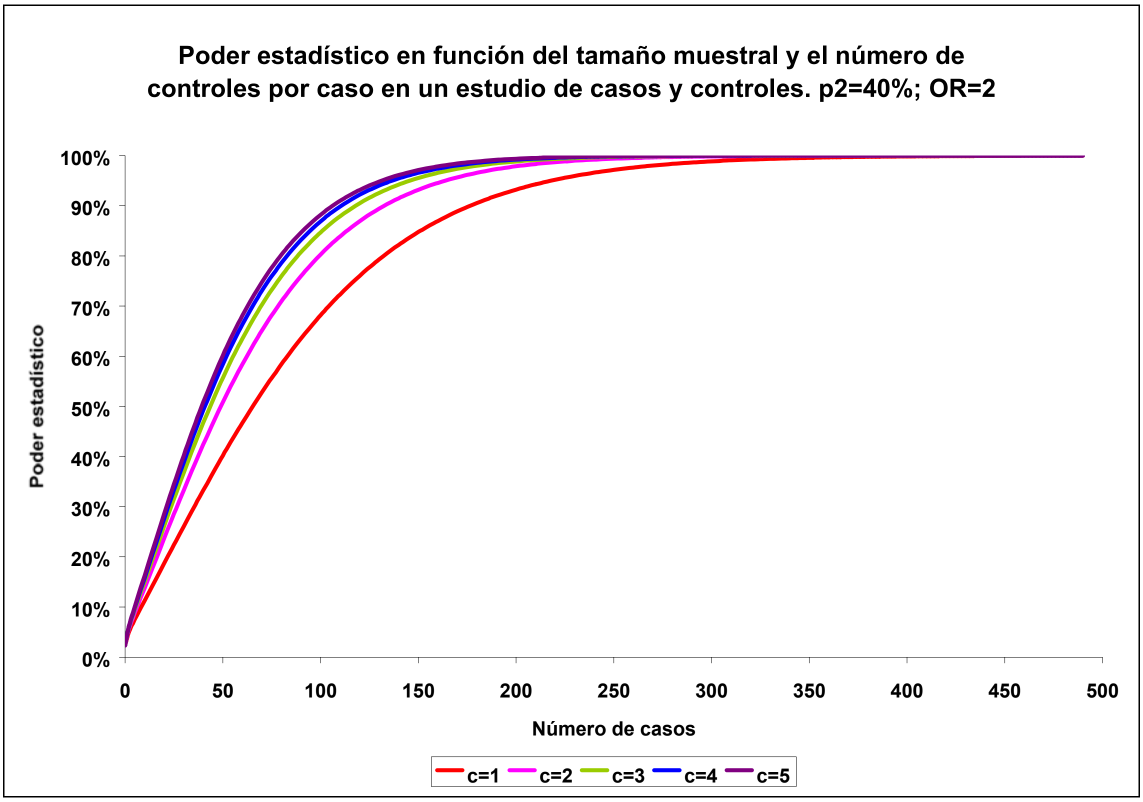

Figura 6.

6. Bibliografía

1. Birnbaum I. Interpreting statistical significance. Teaching Statistics 1982;4(1):24-26.
2. Falk R. Misconceptions of statistical significance. J Struct Learn 1986;9:93-96.
3. Batanero C, Godino JD, Vallecillos A, Green DR, Holmes P. Errors and difficulties in understanding elementary statistical concepts. Int J Math Ed Sci Tech 1994;25(4):527-547.
4. Vallecillos A. Estudio teórico experimental de errores y concepciones sobre el contraste de hipótesis en estudiantes universitarios . Granada: Universidad de Granada; 1994.
5. Vallecillos A. Comprensión de la lógica del contraste de hipótesis en estudiantes universitarios. Rech Didac Math 1995; 15(3):53–81.
6. Batanero C, Vallecillos A. Análisis del aprendizaje de conceptos clave en el contraste de hipótesis estadísticas mediante el estudio de casos. Rech Didac Math 1997; 17(1):29-48.
7. Fisher RA. Statistical methods for research workers. 1ª ed. Edinburg: Oliver & Boyd; 1935.
8. Skipper JK, Guenther, A.S. y Nass, G. The sacredness of .05: a note concerning the uses of statistical levels of significance in social science. Am Sociologist 1967;1:16-18.
9. Gardner MJ, Altman D. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J 1986;292:746-750.

7. Bibliografía adicional

www3.uji.es/~mateu/t6-ig12.doc

www.unizar,es/curso-ice/tomasmar/DocenciaMed.Tra/MT-5web.pdf

www.unizar,es/curso-ice/tomasmar/DocenciaMed.Tra/MT-6web.pdf

www.cead-laspalmas.net/inferencia/cuerpo.htm

www.um.es/estadempresa/estapli2/estapli2.htm
thales.cica.es/rd/Recursos/rd97/UnidadesDidacticas/28-l-u-i.html
cyta.com.ar/biblioteca/bddoc/bdlibros/guia_estadistica/index.htm
www.udc.es/dep/mate/estadistica2/indice_gral.html
www.itch.edu.mx/academic/industrial/estadistca1/toc.html
www.medprev.uma.es/libro/node99.htm
www.sportsci.org/resource/stas/pvalues.html
myphliputil.pearsoncmg/student/levine4/chap06.ppt
myphliputil.pearsoncmg/student/levine4/chap07.ppt
www.ilir.uiuc.edu/courses/lir593/chap71999.ppt